

Joint Channel Selection and Power Control in Infrastructureless Wireless Networks: A Multi-Player Multi-Armed Bandit Framework

Setareh Maghsudi and Sławomir Stańczak, *Senior Member, IEEE*

Abstract

This paper deals with the problem of efficient resource allocation in dynamic infrastructureless wireless networks. Assuming a reactive interference-limited scenario, each transmitter is allowed to select one frequency channel (from a common pool) together with a power level at each transmission trial; hence, for all transmitters, not only the fading gain, but also the number of interfering transmissions and their transmit powers are varying over time. Due to the absence of a central controller and time-varying network characteristics, it is highly inefficient for transmitters to acquire global channel and network knowledge. Therefore a reasonable assumption is that transmitters have no knowledge of fading gains, interference, and network topology. Each transmitting node selfishly aims at maximizing its average reward (or minimizing its average cost), which is a function of the action of that specific transmitter as well as those of all other transmitters. This scenario is modeled as a multi-player multi-armed adversarial bandit game, in which multiple players receive an a priori unknown reward with an arbitrarily time-varying distribution by sequentially pulling an arm, selected from a known and finite set of arms. Since players do not know the arm with the highest average reward in advance, they attempt to minimize their so-called regret, determined by the set of players' actions, while attempting to achieve equilibrium in some sense. To this end, we design in this paper two joint power level and channel selection strategies. We prove that the gap between the average reward achieved by our approaches and that based on the best fixed strategy converges to zero asymptotically. Moreover, the empirical joint frequencies of the game converge to the set of correlated equilibria. We also characterize this set for two special cases

Parts of the material in this paper were presented at the IEEE Wireless Communications and Networking Conference, Shanghai, April, 2013. The work was supported by the German Research Foundation (DFG) under grant STA 864/3-3. The authors are with the Fachgebiet für Informationstheorie und theoretische Informationstechnik, Technische Universität Berlin. The second author is also with the Fraunhofer Institute for Telecommunications Heinrich Hertz Institute, Berlin, Germany (e-mail: setareh.maghsudi@tu-berlin.de, slawomir.stanczak@hhi.fraunhofer.de).

of our designed game. We further discuss experimental regret-testing procedure as another potential solution, which converges to Nash equilibrium. Finally all approaches are compared through extensive numerical analysis.

Index Terms

Adversarial bandits, channel selection, equilibrium, infrastructureless wireless network, power control.

I. INTRODUCTION

A. Bandit Theory and Wireless Communication

Multi-armed bandit (MAB) is a class of sequential optimization problems, to the best of our knowledge originally introduced in [1]. In the most traditional form of MAB, given a set of arms (actions), a player pulls an arm at each trial of the game to receive a reward. The rewards of arms are not known to the player in advance; however, upon pulling an arm, its instantaneous reward is revealed. In such unknown setting, after playing an arm, the player may lose some reward (or incur additional cost) due to not playing another arm instead of the currently played arm. This can be quantified by the difference between the reward that would have been achieved had the player selected another arm, and the reward of the played arm. This quantity is called *regret*. The player decides which arm to pull in a sequence of trials so that its accumulated regret over the game horizon is minimized. Such problems obviously render the intrinsic trade-off between exploration (learning) and exploitation (control), i.e. playing the arm which has exhibited the best performance in the past and playing other arms to guarantee the optimal payoff in future. An important class of bandit games is adversarial bandits, where the series of rewards generated by an arm cannot be attributed to any specific distribution function.

In recent years, bandit theory has been used in communication theory. For instance, [2] and [3] utilize the classical bandit game to model spectrum sharing in cognitive radio networks. In [4], the authors propose a cooperative spectrum sensing scheme based on bandit theory. Further, References [5], [6], and [7] use bandit theory to model relay selection, sensor scheduling and object tracking, respectively. Channel monitoring using bandit model is investigated in [8] and [9]. Bandit models have been also used to solve the distributed resource allocation problem, as discussed in the following.

B. Distributed Resource Allocation in Infrastructureless Wireless Networks

In recent years, game theory and reinforcement learning have been widely used to solve the distributed resource allocation problem. The vast majority of game-theoretic approaches are based on either cooperation (e.g. coalition formation), mechanism design (e.g. auction theory), or exchange economy (e.g. supply-demand markets). Although these approaches can be implemented in a distributed manner, such an implementation in a real network environment requires that each player at least knows its own utility function a priori. On the other hand, these approaches are in general inefficient as players have to exchange information for coordination, which increases signaling and feedback overhead. For example, most models from cooperative game theory require coordination and/or communication among players to construct coalitions [10], [11]. In wireless resource allocation using auction games, bids must be submitted to some central controller that performs necessary computations and makes decisions [12], [13]. Finally, in supply-demand market models, prices and demands are exchanged among buyers and sellers [14], [15].

When the utility functions are not known in advance, the resource allocation problem is often solved by using learning approaches, including bandit models. A large body of literature, such as [16], [17] and [18], analyze single-agent stochastic learning problems. Another example is [19]. In this work, network optimization is modeled as a stochastic bandit game, where at each trial multiple arms are selected by a single player and the reward is some linear combination of the rewards of selected arms. An application of this formulation might be a downlink user selection, performed by the base station. In single-agent settings, the agent learns from its previous experiences, and no information flow is required. However, this type of learning cannot generally be used in wireless networks, where multiple players act selfishly by responding to each other and their utilities are influenced by the actions of other players. Moreover, similar to games with complete information, it is desired that players achieve equilibrium in some sense. As for multi-agent settings, most studies assume that players are able to observe the actions of each other. This assumption, despite being realistic for some spectrum sharing problems, is not always applicable to general resource allocation problems, especially in power control games, where it is difficult to identify the transmit power level of players. In addition, the assumption that each player announces its actions (e.g. its transmit power) is not intensive compatible. As a result, a

great majority of previous works focus on spectrum sharing and/or sensing, as well as channel monitoring. On the other hand, most of previous studies assume that the rewards achieved by each action can be attributed to a single density distribution. However this assumption is highly restrictive especially for dynamic networks.

In [20], multi-agent bandit problem is investigated. This study assumes that in case of interference, no reward is paid to interfering users, thereby eliminating interference, which degrades the overall performance depending on utility functions. In addition, communication among players is necessary. Finally, no equilibrium analysis is performed. Another example is Reference [21], where opportunistic spectrum access is formulated as a multi-agent learning game. In this work, upon availability, each channel pays the same reward to all users so that this scenario is strictly restrictive as it neglects different channel qualities. Moreover, if a channel is selected by multiple users, orthogonal spectrum access scheme is used, which is known to be sub-optimal in general. References [22] and [23] consider graphical games for an interference minimization problem with partially overlapping channels, where the interference is present only between neighboring users. These works establish the convergence of proposed learning approaches for the special case of exact potential games; Nonetheless the analysis does not hold for more general games. The authors of [24] model the cooperative rate maximization in cognitive radio networks as bandit game, and propose two approaches, depending on the availability of information. The stability of the solution is however not investigated. Reference [25] proposes two approaches that achieve Nash equilibrium in a multi-player cognitive environment. System verification, however, is only based on numerical approaches. References [26], [27] and [28] propose various selection schemes to achieve logarithmic regret; however, no equilibrium analysis is performed. All of the works named above assume that the generated rewards of any given action are independent and identically distributed.

C. Our Contribution

As discussed in Section I-B, the resource allocation problem using machine learning theory has been subject to extensive research in recent years. In short, our focus is on a resource allocation problem in an infrastructureless network. First, we model this problem as an adversarial multi-player multi-armed bandit game. With the aim of an efficient management of network resources and the co-channel interference mitigation, we follow an approach suggested in [29] to design

two joint power control and channel selection (PC-CS, hereafter) strategies, which are adapted versions of *exponential-based weighted average* [30] and *follow the leader* [31] strategies. Both PC-CS strategies not only result in small (that is, with sublinear growth rate in time) regret for each individual player, but also guarantee the convergence of empirical frequencies of play to the set of correlated equilibria. We further characterize this set for two special cases of our designed game. Moreover, we implement the *experimental regret-testing procedure* [32], which is shown to converge to the set of Nash equilibria of the game.

Our work extends the state-of-the-art in this area significantly since it differs from the existing studies in the following crucial aspects:

- We analyze the multi-agent bandit problem and take into account the selfishness of players.
- We do not assume that the reward generating process of any given action is time-invariant. In fact, the reward functions are allowed to vary arbitrarily, which enables us to accommodate the dynamic nature of wireless channels and distributed networks.
- We do not allow any communication among players, thereby minimizing the overhead. Moreover, players do *not* observe the actions of each other, so that the developed model can be applied to a large body of resource allocation problems. An example is a power control problem with unknown power levels used by other players. We study a two-dimensional problem, namely joint channel and power level selection problem, by modeling it as a multi-player multi-armed bandit game. In our model, channel qualities are taken into account so that channels pay different rewards to different users. In addition, we impose no limitations on interference pattern.
- Our convergence analysis is valid for a wide range of games. This is in contrast to many previous works where the game should be necessarily potential for the convergence analysis to hold.
- We characterize the set of correlated equilibria for two special cases of our formulated game model.

D. Paper Structure

Section II briefly reviews some concepts and results of bandit theory. In Section III the resource allocation game is formulated. Section IV presents a PC-CS strategy based on *exponential-based weighted average* rule [30]. In Section V, another PC-CS strategy, derived from *follow*

the leader rule [31] is discussed. Section VI is devoted to *experimental regret-testing* procedure [32]. Numerical analysis are presented in Section VII. Section VIII concludes the paper.

II. MULTI-PLAYER MULTI-ARMED BANDIT GAMES

A. Notions of Regret

Multi-player multi-armed bandit problem (MP-MAB, hereafter) is a class of sequential decision making problems with limited information. In this game, each player $k \in \{1, \dots, K\}$ is assigned an action set including N_k actions (arms), $1 \leq N_k \leq N$. Every player selects an action at successive trials in order to receive an initially unknown reward, which is determined not only by its own actions, but also by those of other players. The action set, the played action and the reward achieved by each player are regarded as private information. The reward generating processes of arms are independent. Let \mathbf{I} and $I^{(k)}$ be the joint action space and the action space of player k , respectively. Accordingly, $\mathbf{I}_t = (I_t^{(1)}, \dots, I_t^{(k)}, \dots, I_t^{(K)})$ denotes the joint action profile of players at time t , with $I_t^{(k)}$ being the action of player k . Moreover, let $g_t^{(k)}(\mathbf{I}_t) \in [0, 1]$ be the reward achieved by some player k at time t .¹ The instantaneous regret of any player k is defined as the difference between the reward of the optimal action,² and that of the played action. Based on this definition, the cumulative regret of player k is formally defined in the following.

Definition 1. *The cumulative regret of player k up to time n is defined as*

$$R_n^{(k)} = \max_{i=1, \dots, N_k} \sum_{t=1}^n g_t^{(k)}(i, \mathbf{I}_{t,k}^-) - \sum_{t=1}^n g_t^{(k)}(I_t^{(k)}, \mathbf{I}_{t,k}^-), \quad (1)$$

where $\mathbf{I}_{t,k}^-$ is defined to be the joint action profile of all players except for k at time t .

Each player aims at minimizing its accumulated regret, which is an instance of the well-known exploitation-exploration dilemma: Find a desired balance between exploiting actions that have exhibited well performance in the past (control) on the one hand, and exploring actions which might lead to a better performance in the future (learning) on the other hand.

Now, suppose that players use mixed strategies. This means that, at each trial t , player k selects a probability distribution $\mathbf{P}_t^{(k)} = (p_{1,t}^{(k)}, \dots, p_{i,t}^{(k)}, \dots, p_{N_k,t}^{(k)})$ over arms, and plays arm i with

¹Note that all results can be also expressed in terms of loss (d), provided that the loss is related to the gain by $d = 1 - g$, $g \in [0, 1]$.

²Optimality is defined in the sense of the highest instantaneous reward.

probability $p_{i,t}^{(k)}$. In this case, we resort to expected regret, also called *external regret* [33], defined as follows.

Definition 2. *The external cumulative regret of player k is defined as*

$$\begin{aligned} R_{\text{Ext}}^{(k)} &:= R_{\text{Ext}}^{(k)}(n) = \max_{i=1,\dots,N_k} \sum_{t=1}^n g_t^{(k)}(i, \mathbf{I}_{t,k}^-) - \sum_{t=1}^n \bar{g}_t^{(k)}(\mathbf{P}_t^{(k)}, \mathbf{I}_{t,k}^-) \\ &= \max_{i=1,\dots,N_k} \sum_{t=1}^n \sum_{j=1}^N p_{j,t}^{(k)} \left(g_t^{(k)}(i, \mathbf{I}_{t,k}^-) - g_t^{(k)}(j, \mathbf{I}_{t,k}^-) \right), \end{aligned} \quad (2)$$

where $\bar{g}_t^{(k)}(\cdot)$ denotes the expected reward at round t by using mixed strategy $\mathbf{P}_t^{(k)}$, defined as $\bar{g}_t^{(k)}(\cdot) = \sum_{j=1}^N g_t^{(k)}(\cdot) p_{j,t}^{(k)}$.

By definition, external regret compares the expected reward of the current mixed strategy with that of the best fixed action in the hindsight, but fails to compare the rewards achieved by changing actions in a pair-wise manner. In order to compare actions in pairs, *internal regret* [33] is introduced that is closely related to the concept of equilibrium in games.

Definition 3. *The internal cumulative regret of player k is defined as*

$$\begin{aligned} R_{\text{Int}}^{(k)} &:= R_{\text{Int}}^{(k)}(n) = \max_{i,j=1,\dots,N_k} R_{(i \rightarrow j),n}^{(k)} \\ &= \max_{i,j=1,\dots,N_k} \sum_{t=1}^n p_{i,t}^{(k)} \left(g_t^{(k)}(j, \mathbf{I}_{t,k}^-) - g_t^{(k)}(i, \mathbf{I}_{t,k}^-) \right). \end{aligned} \quad (3)$$

Notice that on the right-hand side of (3), $r_{(i \rightarrow j),t}^{(k)} = p_{i,t}^{(k)} \left(g_t^{(k)}(j, \cdot) - g_t^{(k)}(i, \cdot) \right)$ denotes the expected regret caused by pulling arm i instead of arm j . By comparing (2) and (3), external regret can be bounded above by internal regret as [34]

$$R_{\text{Ext}}^{(k)} = \max_{i=1,\dots,N_k} \sum_{j=1}^{N_k} R_{(i \rightarrow j),n}^{(k)} \leq N_k \max_{i,j=1,\dots,N_k} R_{(i \rightarrow j),n}^{(k)} = N_k R_{\text{Int}}^{(k)}. \quad (4)$$

Remark 1. *Throughout the paper, vanishing (zero-average) external and internal regret means that $\lim_{n \rightarrow \infty} \frac{1}{n} R_{\text{Ext}} = 0$ and $\lim_{n \rightarrow \infty} \frac{1}{n} R_{\text{Int}} = 0$, respectively. In other words, we have $R_{\text{Ext}} \in o(n)$ and $R_{\text{Int}} \in o(n)$. Note that by (4), $R_{\text{Int}} \in o(n)$ yields $R_{\text{Ext}} \in o(n)$. Throughout the paper, we call any strategy with $R_{\text{Int}} \in o(n)$ as "no-regret strategy".*

B. Equilibrium

From the view point of each player k , an MP-MAB is seen as a game with two agents: player k itself, and the *set* of all other $K - 1$ players (referred to as the opponent), whose joint action profile affects the reward achieved by player k . We consider here the most general framework, where the opponent is non-oblivious, i.e. its series of actions depends on the actions of player k . It is known that a game against a non-oblivious opponent can be modeled *only* by adversarial bandit games [35], while similar to other game-theoretic formulations, the solution is considered to be equilibrium, most importantly Nash and correlated equilibria.³

In the context of game-theoretic bandits, an important result is the following theorem.

Theorem 1 ([33]). *Consider a K -player bandit game, where each player k is provided with an action set of cardinality N_k . Denote the internal regret of player k by $R_{\text{Int}}^{(k)}$, and the set of correlated equilibria by \mathfrak{C} . At time n , define the empirical joint distribution of the game as*

$$\hat{\pi}_n(\mathbf{i}) = \frac{1}{n} \sum_{t=1}^n \mathbb{I}_{\{\mathbf{I}_t=\mathbf{i}\}}, \quad \mathbf{i} = (i^{(1)}, \dots, i^{(K)}) \in \bigotimes_{k=1}^K \{1, \dots, N_K\}. \quad (5)$$

Then, if all players $k \in \{1, \dots, K\}$ play according to any strategy so that

$$\lim_{n \rightarrow \infty} \frac{1}{n} R_{\text{Int}}^{(k)} = 0, \quad (6)$$

the distance $\inf_{\pi \in \mathfrak{C}} \sum_i |\hat{\pi}_n(\mathbf{i}) - \pi(\mathbf{i})|$ between the empirical joint distribution of plays and the set of correlated equilibria converges to 0 almost surely.

Theorem 1 simply states that in an MP-MAB game, if all players play according to a strategy with vanishing internal regret (no-regret), then the empirical joint distribution of plays converges to the set of correlated equilibria. Note that the strategies used by players are not required to be identical. Since a rational player is always interested in minimizing its regret, the assumption that every player plays according to a no-regret strategy is reasonable.

C. From Vanishing External Regret to Vanishing Internal Regret

In [34], an approach is proposed for converting any selection strategy with vanishing external regret to another version with vanishing internal regret. We describe this approach briefly.

³These definitions are quite standard (see e.g. [36]), and thus we do not restate them here.

Consider a selection strategy (O-strategy, hereafter) which at each time t assigns probability distribution \mathbf{P}_t to the set of N actions, and selects an action according to this distribution. Assume that the player starts using O-Strategy with uniform distribution over N actions. At each time $t > 1$, the O-strategy has already selected $\mathbf{P}_{t-1} = (p_{1,t-1}, \dots, p_{i,t-1}, \dots, p_{j,t-1}, \dots, p_{N,t-1})$. Now, the O-strategy constructs a meta-strategy (M-strategy, hereafter) with $N(N-1)$ virtual strategies based on \mathbf{P}_{t-1} . Each virtual strategy corresponds to a pair of actions $(i \rightarrow j)$, $(i, j \in \{1, \dots, N\}, i \neq j)$, and constructs a distribution over N actions by assigning the probability mass of action i to action j . That is, it defines $\mathbf{P}_{t-1}^{(i \rightarrow j)} = (p_{1,t-1}, \dots, 0, \dots, p_{j,t-1} + p_{i,t-1}, \dots, p_{N,t-1})$, which has 0 and $p_{j,t-1} + p_{i,t-1}$ at the place of $p_{i,t-1}$ and $p_{j,t-1}$, respectively, and all other elements remain unchanged. Assume that the M-strategy treats these virtual strategies as actions. That is, at each time t , it defines a probability vector δ_t over $N(N-1)$ virtual actions, where the probability of action $(i \rightarrow j)$, i.e. $\delta_{(i \rightarrow j),t}$, depends on its past performance.⁴ Now, at time t , the O-strategy assigns a distribution \mathbf{P}_t to N actions, where $\mathbf{P}_t = \sum_{(i,j): i \neq j} \mathbf{P}_t^{(i \rightarrow j)} \delta_{(i \rightarrow j),t}$. The constructed O-strategy has the characteristic that its internal regret is upper-bounded by the external regret of the M-strategy over $N(N-1)$ virtual actions according to probability δ_t . Thus, if the M-strategy exhibits vanishing external regret, the O-strategy results in vanishing internal regret. In Section IV and V, we use this property to design no-regret selection strategies.

III. BANDIT-THEORETICAL MODEL OF INFRASTRUCTURELESS WIRELESS NETWORKS

We consider a network consisting of K transmitter-receiver pairs, denoted by (k, k') , where $k, k' \in \{1, \dots, K\}$. The transmitter-receiver pair (k, k') is referred to as user or player k . Each user k can access C_k mutually orthogonal channels at L_k quantized power levels. This implies that its strategy set includes $N_k = C_k \times L_k$ actions, where at time t each action $I_t^{(k)} = (c_t^{(k)}, l_t^{(k)})$ consists of one channel index (which corresponds to some channel quality), and one power level. Therefore, the joint action profile of users, \mathbf{I}_t , is to be understood here as the pair $(\mathbf{c}_t, \mathbf{l}_t)$, where $\mathbf{c}_t = (c_t^{(1)}, \dots, c_t^{(K)})$ and $\mathbf{l}_t = (l_t^{(1)}, \dots, l_t^{(K)})$. As each channel might be accessible by multiple users, co-channel interference (collision, interchangeably) is likely to arise. Since users are allowed to select a new channel and to adapt their power levels at each transmission trial, interference pattern in general changes over time. In addition, the distribution of fading coefficients might be

⁴Note that the gains of virtual actions cannot be calculated explicitly. Later we will see that the gain achieved by any virtual action $(i \rightarrow j)$ is calculated based on the gain achieved by playing true actions i and j .

time-varying so that acquiring channel and/or network information at the level of autonomous transmitters would be extremely challenging and inefficient. Therefore, we assume that

- (A1) transmitters have *no* channel knowledge or any other side information such as the number of users or their selected actions.
- (A2) In addition, users do not coordinate their actions that can be chosen completely asynchronously by each user.

Note that as users do not observe the actions of each other, it might be in their interest to select their actions at the beginning of trials, thereby using the remaining time for data transmission.

In this paper, we model the joint channel and power level selection problem as a K -player adversarial bandit game, where player k decides for one of the N_k actions. We define the expected utility function (reward) of player k to be⁵

$$G_t^{(k)}(\mathbf{I}) = \log \left(\frac{l^{(k)} |h_{kk',t,c^{(k)}}|^2}{\sum_{q=1}^{Q_k} l^{(q)} |h_{qk',t,c^{(k)}}|^2 + N_0} \right) - \alpha \cdot l^{(k)}, \quad (7)$$

for some given joint action profile $\mathbf{I} = (\mathbf{c}, \mathbf{l})$. In (7), $Q_k < K$ is the number of players that interfere with user k in channel $c^{(k)}$. Throughout the paper, $|h_{uv,t,c}|^2 \in \mathbb{R}^+$ is used to denote the average gain of channel c between $u \rightarrow v$ at time t . N_0 is the variance of zero-mean additive white Gaussian noise, and $\alpha \geq 0$ is the constant power price factor. The last term in (7) is used to penalize the use of excessive power. According to Section II, let $g_t^{(k)}(\mathbf{I}_t) \in [0, 1]$ denote the achieved reward of player k at time t , as a function of joint action profile \mathbf{I}_t . We consider a game with noisy rewards where $g_t^{(k)}(\mathbf{I}) = G_t^{(k)}(\mathbf{I}) + \epsilon_t$, with ϵ being some zero-mean random variable with bounded variance, which is independent and identically distributed over time. As it is well-known, in a non-cooperative game, the primary goal of each selfish player is to maximize its own accumulated reward. Formally, this can be written as

$$\text{maximize}_{(c_t^{(k)}, l_t^{(k)})} \sum_{t=1}^n g_t^{(k)}(\mathbf{c}_t, \mathbf{l}_t), \quad (8)$$

where $c_t^{(k)} \in \{1, \dots, C_k\}$ and $l_t^{(k)} \in \{1, \dots, L_k\}$. By Assumptions (A1) and (A2), however, it is clear that the objective function in (8) is not available. For this reason, we argue for a less ambitious goal, which is known as *regret minimization*. More precisely, each player k attempts

⁵Throughout the paper, logarithms are based 2 unless otherwise is stated.

to achieve vanishing external regret in the sense that

$$\begin{aligned} & \lim_{n \rightarrow \infty} \frac{1}{n} R_{\text{Ext}}^{(k)} \\ &= \lim_{n \rightarrow \infty} \frac{1}{n} \left(\max_{i=1, \dots, N_k} \sum_{t=1}^n g_t^{(k)}(i, \mathbf{I}_{t,k}^-) - \sum_{t=1}^n \bar{g}_t^{(k)}(\mathbf{P}_t^{(k)}, \mathbf{I}_{t,k}^-) \right) = 0. \end{aligned} \quad (9)$$

In addition to the individual strategy of each user aiming at satisfying (9), all players should achieve some steady state, i.e. *equilibrium*. Therefore, in the remainder of this paper, we develop algorithmic solutions to the resource allocation problem with a twofold objective in mind: i) external regret of each user should vanish asymptotically according to (9) and ii) the actions of all players should convergence to equilibrium.

By (4), the external regret of each user is upper-bounded by its internal regret. As a result, if all users select their actions according to some no-regret strategy, not only (9) is achieved by all of them (see also Remark 1), but also the corresponding game converges to equilibrium in some sense, which immediately follows from Theorem 1. In Sections IV and V, we present two internal-regret minimizing strategies that are shown to solve the game and, with it, to achieve the two objectives mentioned above. Both algorithms can be applied in a *fully decentralized* manner by each player, since at each time, they only require the set of past rewards of the respective player.

Finally, it is worth noting that the set of correlated equilibria for the general time-varying repeated game defined by (7) cannot be characterized. Nevertheless, in what follows, we characterize this set for two games defined by some relaxed versions of (7). First, consider a game similar to the one defined above, with the difference that unlike (7), the reward process is assumed to be stationary, i.e.

$$G^{(k)}(\mathbf{I}) = \log \left(\frac{l^{(k)} |h_{kk', c^{(k)}}|^2}{\sum_{q=1}^{Q_k} l^{(q)} |h_{qk', c^{(k)}}|^2 + N_0} \right) - \alpha \cdot l^{(k)}, \quad (10)$$

which implies that the average channel gains are time-invariant. By the following proposition, this game has a unique correlated equilibrium.

Proposition 1. *Consider a K -player game where the expected reward function of each player k is defined by (10). This game has a unique correlated equilibrium which places probability one on its unique pure-strategy Nash equilibrium.*

Proof: See Section IX-B. ■

Now let the expected reward function be defined as follows:

$$G^{(k)}(\mathbf{I}) = \log\left(l^{(k)} \frac{|h_{kk',c^{(k)}}|^2}{N_0}\right) - \alpha l^{(k)}, \quad (11)$$

which is more restricted, but simpler than (10). With this choice of expected reward function, the game can be shown to have a unique correlated equilibrium that maximizes the aggregate utility of all players, i.e. the social welfare. This result is stated formally in the following proposition.

Proposition 2. *Consider a K -player game where each player k has the expected reward function $G^{(k)}(\cdot)$ given by (11). This game has a unique correlated equilibrium which places probability one on a unique pure strategy Nash equilibrium that maximizes $\sum_{k=1}^K G^{(k)}(\cdot)$.*

Proof: See Section IX-C. ■

IV. NO-REGRET BANDIT EXPONENTIAL-BASED WEIGHTED AVERAGE STRATEGY

The basic idea of an exponential-based weighted strategy is to assign each action, at every trial, some selection probability which is inversely proportional to exponentially-weighted accumulated regret (or directly proportional to exponentially-weighted accumulated reward) caused by that action in the past [37]. Roughly speaking, if playing an action has resulted in large regret in the past, its future selection probability is small, and vice versa.

As described in Section II-A, in bandit formulation, players only observe the reward of the played action, and not those of others. Therefore the reward of each action i is estimated as [33]

$$\tilde{g}_t^{(k)}(i) = \begin{cases} \frac{g_t^{(k)}(I_t^{(k)})}{p_{i,t}^{(k)}} & i = I_t^{(k)} \\ 0 & o.w. \end{cases}, \quad (12)$$

which is an unbiased estimate of the true reward of action i ; that is, $\mathbb{E}_t[\tilde{g}^{(k)}(i)] = g^{(k)}(i)$. Estimated rewards are afterwards used to calculate regrets. For example, the regret of *not* playing action j instead of action i yields

$$\tilde{R}_{(i \rightarrow j), t-1}^{(k)} = \sum_{s=1}^{t-1} \tilde{r}_{(i \rightarrow j), s}^{(k)} = \sum_{s=1}^{t-1} p_{i,s}^{(k)} (\tilde{g}_s^{(k)}(j) - \tilde{g}_s^{(k)}(i)). \quad (13)$$

Despite exhibiting vanishing external regret, weighted average strategies yield in general large internal regret; as a result, even if all players play according to such strategies, the game does *not*

converge to equilibrium. In the following, we utilize the bandit version of exponentially weighted average strategy [38], and convert it to an improved version that yields small *internal* regret, using the approach of Section II-C. The strategy is called no regret bandit exponentially-weighted average strategy (NR-BEWAS), and is described in Algorithm 1.

Algorithm 1 No-Regret Bandit Exponential-Based Weighted Average Strategy (NR-BEWAS)

- 1: If the game horizon, n , is known, define γ_t and η_t as given in Proposition 3, otherwise as those given in Proposition 4.
 - 2: Define $\Phi(\mathbf{U}) = \frac{1}{\eta_t} \ln \left(\sum_{i=1}^{N_k} \exp(\eta_t u_i) \right)$, where $\mathbf{U} = (u_1, \dots, u_{N_k}) \in \mathbb{R}^{N_k}$.
 - 3: Let $\mathbf{P}_1^{(k)} = \left(\frac{1}{N_k}, \dots, \frac{1}{N_k} \right)$ (uniform distribution).
 - 4: Select an action using $\mathbf{P}_1^{(k)}$.
 - 5: Play and observe the reward.
 - 6: **for** $t = 2, \dots, n$ **do**
 - 7: Let $\mathbf{P}_{t-1}^{(k)}$ be the mixed strategy at time $t-1$, i.e. $\mathbf{P}_{t-1}^{(k)} = (p_{1,t-1}^{(k)}, \dots, p_{i,t-1}^{(k)}, \dots, p_{j,t-1}^{(k)}, \dots, p_{N_k,t-1}^{(k)})$.
 - 8: Construct $\mathbf{P}_{t-1}^{(k),(i \rightarrow j)}$ as follows: replace $p_{i,t-1}^{(k)}$ in $\mathbf{P}_{t-1}^{(k)}$ by zero, and instead increase $p_{j,t-1}^{(k)}$ to $p_{j,t-1}^{(k)} + p_{i,t-1}^{(k)}$. Other elements remain unchanged. We obtain $\mathbf{P}_{t-1}^{(k),(i \rightarrow j)} = (p_{1,t-1}^{(k)}, \dots, 0, \dots, p_{j,t-1}^{(k)} + p_{i,t-1}^{(k)}, \dots, p_{N_k,t-1}^{(k)})$.
 - 9: Define

$$\delta_{(i \rightarrow j),t}^{(k)} = \frac{\exp \left(\eta_t \tilde{R}_{(i \rightarrow j),t-1}^{(k)} \right)}{\sum_{(m \rightarrow l): m \neq i} \exp \left(\eta_t \tilde{R}_{(m \rightarrow l),t-1}^{(k)} \right)}, \quad (14)$$
 - 10: where $\tilde{R}_{(i \rightarrow j),t-1}^{(k)}$ is calculated by using (12) and (13).
 Given $\delta_{(i \rightarrow j),t}^{(k)}$, solve the following fixed point equation to find $\mathbf{P}_t^{(k)}$:

$$\mathbf{P}_t^{(k)} = \sum_{(i \rightarrow j): i \neq j} \mathbf{P}_t^{(k),(i \rightarrow j)} \delta_{(i \rightarrow j),t}^{(k)}. \quad (15)$$
 - 11: Final probability distribution yields

$$\mathbf{P}_t^{(k)} = (1 - \gamma_t) \mathbf{P}_t^{(k)} + \frac{\gamma_t}{N_k}. \quad (16)$$
 - 12: Using the final $\mathbf{P}_t^{(k)}$, given by (16), select an action.
 - 13: Play and observe the reward.
 - 14: **end for**
-

From Algorithm 1, NR-BEWAS has two parameters, namely γ_t and η_t . In the event that the game horizon, n , is known in advance, these two parameters are constant over time ($\eta_t = \eta$ and $\gamma_t = \gamma$), and the growth rate of regret can be bounded precisely, mainly based on the results of [33]. Otherwise, they vary with time. In this case, vanishing (sub-linear in time) internal regret can be guaranteed; nevertheless, this bound might be loose. This discussion is formalized by following propositions.

Proposition 3. Let $\eta_t = \eta = \left(\frac{\ln N_k}{2N_k n}\right)^{\frac{2}{3}}$ and $\gamma_t = \gamma = \left(\frac{N_k^2 \ln N_k}{4n}\right)^{\frac{1}{3}}$. Then Algorithm 1 (NR-BEWAS) yields vanishing internal regret and we have $R_{\text{int}}^{(k)} \in O((nN_k^2 \ln N_k)^{\frac{2}{3}})$.

Proof: See Appendix IX-D. ■

Proposition 4. Let $\eta_t = \frac{\gamma_t^3}{N_k^2}$ and $\gamma_t = t^{-\frac{1}{3}}$. Then Algorithm 1 (NR-BEWAS) yields vanishing internal regret; that is we have $R_{\text{int}}^{(k)} \in o(n)$.

Proof: See Appendix IX-E. ■

The following corollaries follow from the above propositions and Theorem 1.

Corollary 1. If all players play according to NR-BEWAS, then the empirical joint frequencies of play converge to the set of correlated equilibria.

Proof: The proof is a direct consequence of Theorem 1 and Proposition 3 or Proposition 4. ■

Corollary 2. Let ϵ -correlated equilibrium approximate correlated equilibrium in the sense that $\bigcap_{\epsilon>0} \mathfrak{C}_\epsilon = \mathfrak{C}$. Assuming that the game horizon is known and all players play according to NR-BEWAS, then the minimum required number of trials to achieve ϵ -correlated equilibrium yields $\max_{k=1,\dots,K} \epsilon^{-\frac{3}{2}} O((N_k K)(N_k^2 \ln N_k + K^2 \ln K))$, which is proportional to $\epsilon^{-\frac{3}{2}}$ and increases polynomially in the number of actions as well as in the number of players.

Proof: The proof follows from the bound of Proposition 3 and Remark 7.6 of [33].⁶ ■

V. NO-REGRET BANDIT FOLLOW THE PERTURBED LEADER STRATEGY

Similar to the weighted-average strategy presented in the previous section, the strategy *follow the perturbed leader* is an approach to solve online decision-making problems. In the basic version of this approach, called *follow the leader* [39], the action with the minimum regret in the past is selected at each trial. However, this method is deterministic and therefore does not achieve vanishing regret against non-oblivious opponents. Therefore, in *follow the perturbed leader*, player adds a random perturbation to the vector of accumulated regrets, and the action with the minimum perturbed regret in the past is selected [33]. In [40], a bandit version of this

⁶Details are omitted to avoid unnecessary restatement of existing analysis.

algorithm is constructed, where unobserved rewards are estimated. The authors show that the developed algorithm exhibits vanishing *external regret*. Similar to NR-BEWAS, we here modify the algorithm of [40] to ensure vanishing *internal regret*. The approach is called no-regret bandit follow the perturbed leader strategy (NR-BFPLS).

Algorithm 2 No-Regret Bandit Follow the Perturbed Leader Strategy (NR-BFPLS)

- 1: Define $\epsilon_t = \epsilon_n = \frac{\sqrt{\ln n}}{3\sqrt{N_k n}}$, and $\gamma_t = \min(1, N_k \epsilon_t)$. Note that unlike NR-BEWAS, here we know the game horizon (n) in advance.
- 2: Let $\mathbf{P}_1^{(k)} = \left(\frac{1}{N_k}, \dots, \frac{1}{N_k}\right)$ (uniform distribution).
- 3: Select an action using $\mathbf{P}_1^{(k)}$.
- 4: Play and observe the reward.
- 5: **for** $t = 2, \dots, n$ **do**
- 6: Let $\mathbf{P}_{t-1}^{(k)}$ be the mixed strategy at time $t - 1$, i.e. $\mathbf{P}_{t-1}^{(k)} = (p_{1,t-1}^{(k)}, \dots, p_{i,t-1}^{(k)}, \dots, p_{j,t-1}^{(k)}, \dots, p_{N_k,t-1}^{(k)})$.
- 7: Construct $\mathbf{P}_{t-1}^{(k),(i \rightarrow j)}$ as follows: replace $p_{i,t-1}^{(k)}$ in $\mathbf{P}_{t-1}^{(k)}$ by zero, and instead increase $p_{j,t-1}^{(k)}$ to $p_{j,t-1}^{(k)} + p_{i,t-1}^{(k)}$. Other elements remain unchanged. We obtain $\mathbf{P}_{t-1}^{(k),(i \rightarrow j)} = (p_{1,t-1}^{(k)}, \dots, 0, \dots, p_{j,t-1}^{(k)} + p_{i,t-1}^{(k)}, \dots, p_{N_k,t-1}^{(k)})$.
- 8: Calculate $\tilde{R}_{(i \rightarrow j),t-1}^{(k)}$ using (12) and (13).
- 9: Define $\sigma_{(i \rightarrow j),t-1} = \left(\sum_{\tau=1}^{t-1} \frac{1}{\delta_{(i \rightarrow j),\tau}^{(k)}}\right)^{\frac{1}{2}}$, which is the upper-bound of conditional variances of random variables $\tilde{R}_{(i \rightarrow j),t-1}^{(k)}$ [40].
- 10: Let $\tilde{R}_{(i \rightarrow j),t-1}^{(k)} = \tilde{R}_{(i \rightarrow j),t-1}^{(k)} - \sqrt{1 + \sqrt{2/N_k} \sigma_{(i \rightarrow j),t-1} \sqrt{\ln(t)}}$ [40].
- 11: Randomly select a perturbation vector $\underline{\mu}_t$ with $N_k(N_k - 1)$ elements from two-sided exponential distribution with width ϵ_t .
- 12: Consider a selection rule which selects the action $(i \rightarrow j)$ given by

$$\operatorname{argmax} \left\{ \tilde{R}_{(i \rightarrow j),t-1}^{(k)} + \mu_{(i \rightarrow j),t} \right\}, \quad (i \rightarrow j) \in \{1, \dots, N_k(N_k - 1)\} \quad (17)$$

Note that in our setting $\tilde{R}_{(i \rightarrow j)}$ denotes the estimated regret of *not* playing action $(i \rightarrow j)$, hence we find the action with largest \tilde{R} .

- 13: From (17), calculate the probability $\delta_{(i \rightarrow j),t}^{(k)}$ assigned to each pair $(i \rightarrow j)$.
- 14: Given $\delta_{(i \rightarrow j),t}^{(k)}$, solve the following fixed point equation to find $\mathbf{P}_t^{(k)}$.

$$\mathbf{P}_t^{(k)} = \sum_{(i \rightarrow j): i \neq j} \mathbf{P}_t^{(k),(i \rightarrow j)} \delta_{(i \rightarrow j),t}^{(k)}. \quad (18)$$

- 15: Final probability distribution yields

$$\mathbf{P}_t^{(k)} = (1 - \gamma_t) \mathbf{P}_t^{(k)} + \frac{\gamma_t}{N_k}. \quad (19)$$

- 16: Using the final $\mathbf{P}_t^{(k)}$, given by (19), select an action.
 - 17: Play and observe the reward.
 - 18: **end for**
-

Algorithm 2 requires the knowledge of the probability assigned to each action by the *follow the perturbed leader* strategy at every trial. However, in contrast to NR-BEWAS, these probabilities

are not assigned explicitly; therefore we explain how to calculate these values.

From (17), the selection probability of virtual action $(i \rightarrow j) \in \{1, \dots, N_k(N_k - 1)\}$ is the probability that $\tilde{R}_{(i \rightarrow j), t-1}$ plus perturbation $\mu_{(i \rightarrow j), t}$ is larger than those of other actions, i.e.

$$\begin{aligned}
& \Pr[I_t = (i \rightarrow j)] \\
&= \Pr[\tilde{R}_{(i \rightarrow j), t-1} + \mu_{(i \rightarrow j), t} \geq \tilde{R}_{(i' \rightarrow j'), t-1} + \mu_{(i' \rightarrow j'), t} \quad \forall (i \rightarrow j) \neq (i' \rightarrow j')] \\
&= \int_{-\infty}^{\infty} \Pr[\tilde{R}_{(i \rightarrow j), t-1} + \mu_{(i \rightarrow j), t} = m \wedge \tilde{R}_{(i' \rightarrow j'), t-1} + \mu_{(i' \rightarrow j'), t} \leq m \quad \forall (i \rightarrow j) \neq (i' \rightarrow j')] dm \\
&= \int_{-\infty}^{\infty} \Pr[\tilde{R}_{(i \rightarrow j), t-1} + \mu_{(i \rightarrow j), t} = m] \prod_{(i' \rightarrow j') \neq (i \rightarrow j)} \Pr[\tilde{R}_{(i' \rightarrow j'), t-1} + \mu_{(i' \rightarrow j'), t} \leq m] dm.
\end{aligned} \tag{20}$$

Since μ_t is distributed according to a two-sided exponential distribution with width ϵ_n , the terms under integral can be calculated easily (see [41], for example). Now we are in a position to show some properties of NR-BFPLS (Algorithm 2).

Proposition 5. *Let $\epsilon_t = \epsilon = \frac{\sqrt{\ln n}}{3\sqrt{N_k n}}$ and $\gamma_t = \gamma = \min(1, N_k \epsilon_t)$. Then Algorithm 2 (NR-BFPL) yields vanishing internal regret with $R_{\text{Int}}^{(k)} \in O((nN_k^2 \ln N_k)^{\frac{1}{2}})$.*

Proof: By [40], we know that if the BPFL algorithm is applied to N_k actions, then $R_{\text{Ext}}^{(k)} \in O((nN_k \ln N_k)^{\frac{1}{2}})$. Using this, the proof proceeds along similar lines as the proof of Proposition 3 and is therefore omitted here. ■

Corollary 3. *Assuming that the game horizon is known and all players play according to NR-BFPLS, then the minimum required number of trials to achieve ϵ -correlated equilibrium yields $\max_{k=1, \dots, K} \epsilon^{-2} O((N_k K)(N_k^2 \ln N_k + K^2 \ln K))$, which is proportional to ϵ^{-2} and increases polynomially in the number of actions as well as in the number of players.*

Proof: The proof is a result of the bound of Proposition 5 and Remark 7.6 of [33]. ■

VI. BANDIT EXPERIMENTAL REGRET-TESTING STRATEGY

Experimental regret-testing belongs to the large family of exhaustive search algorithms, and is comprehensively discussed in [32] and [33] for bandit games. In this section, we briefly review this approach, and investigate its performance later in Section VII-A.

First, the time is divided into periods $m = 1, 2, \dots$ of length T so that for each m we have

$t \in [(m-1)T+1, mT]$. At the beginning of period m , any player k randomly selects a mixed strategy, denoted by $\mathbf{P}_m^{(k)}$. Moreover, some random variable $U_{k,t}^{(m)} \in \{1, \dots, n_k, \dots, N_k\}$ is defined as follows. For $t \in [(m-1)T+1, mT]$, and for each n_k , there are exactly s values of t such that $U_{k,t}^{(m)} = n_k$, and $U_{k,t}^{(m)} = 0$ for the remaining $t = T - sN_k$ trials. At time t , the action $I_t^{(k)}$ is selected to be [38]

$$I_t^{(k)} : \begin{cases} \text{is distributed as } \mathbf{P}_m^{(k)} & \text{if } U_{k,t}^{(m)} = 0 \\ \text{equals } n_k & \text{if } U_{k,t}^{(m)} = n_k \end{cases}. \quad (21)$$

At the end of period m , player k calculates the experimental regret of playing each action n_k as [38]

$$\hat{r}_{m,n_k}^{(k)} = \frac{1}{T - sN_k} \sum_{t=(m-1)T+1}^{mT} g_t^{(k)}(\mathbf{I}_t) \mathbb{I}_{\{U_{k,t}^{(m)}=0\}} - \frac{1}{s} \sum_{t=(m-1)T+1}^{mT} g_t^{(k)}(n_k, \mathbf{I}_{t,k}^-) \mathbb{I}_{\{U_{k,t}^{(m)}=n_k\}}. \quad (22)$$

If the regret is smaller than an acceptable threshold ρ , the player continues to play its current mixed strategy. Otherwise, another mixed strategy is selected. The procedure is summarized in Algorithm 3. It is known that if the parameters of BERTS (e.g. T and ρ) are chosen appropriately, then, in a long run, the played mixed strategy profile is an approximate Nash equilibrium for almost all the time. Details can be found in [33], and hence are omitted.

Algorithm 3 Bandit Experimental Regret Testing Strategy [33] (BERTS)

- 1: Set T (period length), ρ (acceptable regret threshold), $\xi \ll 1$ (exploration parameter), $m = 1$ (period index). Notice that for each period $m = 1, \dots, M$, we have $t \in [(m-1)T+1, mT]$.
 - 2: Select a mixed strategy, $\mathbf{P}_m^{(k)}$ according to the uniform distribution, from the probability simplex with N_k dimensions.
 - 3: For each $n_k \in \{1, \dots, N_k\}$ select s exploring trials at random. Exploration trials which are dedicated to different actions should not overlap.
 - 4: **for** $t = (m-1)T + y$, where $1 \leq y < T$ **do**
 - 5: **if** t is an exploring trial dedicated to action i **then**
 - 6: play action i and observe the reward.
 - 7: **else**
 - 8: select an action using $\mathbf{P}_m^{(k)}$. Play and observe the reward.
 - 9: **end if**
 - 10: **end for**
 - 11: Calculate the experimental regret of period m , $\hat{r}_{m,n_k}^{(k)}$, using (22);
 - 12: **if** $\max_{n_k=1, \dots, N_k} \hat{r}_{m,n_k}^{(k)} > \rho$, **then**
 - 13: 1) set $m = m + 1$, 2) go to line 2.
 - 14: **else**
 - 15: • with probability ξ : 1) set $m = m + 1$, 2) go to line 2;
 - 15: • with probability $1 - \xi$: 1) let $\mathbf{P}_{m+1}^{(k)} = \mathbf{P}_m^{(k)}$, 2) set $m = m + 1$, 3) go to line 3.
 - 16: **end if**
-

VII. NUMERICAL ANALYSIS

Numerical analysis consists of two parts. In Section VII-A, we consider a simple network, and clarify the work flow of algorithms. In Section VII-B, we consider a larger network, and study the performance of the proposed game model and algorithmic solutions in comparison with some other selection strategies.

A. Part One

1) *Network model*: The network consists of two transmitter-receiver pairs (users). There exist two orthogonal channels, C_1 and C_2 , and two power-levels, P_1 and P_2 . Hence, the action set of each user yields $\{a_1 : (C_1, P_1), a_2 : (C_1, P_2), a_3 : (C_2, P_1), a_4 : (C_2, P_2)\}$. The distribution of channel gains changes at each trial. We assume that the variance of mean values of these distributions is relatively small, which corresponds to *low dynamicity*.⁷ Channel matrices are $H_1 = \begin{bmatrix} [0.50, 0.80] & [0.15, 0.20] \\ [0.01, 0.05] & [0.01, 0.09] \end{bmatrix}$ and $H_2 = \begin{bmatrix} [0.02, 0.05] & [0.02, 0.06] \\ [0.05, 0.15] & [0.75, 0.95] \end{bmatrix}$, where $H_{l,(u,v)}$ ($u, v, l \in \{1, 2\}$), corresponds to the link $u \rightarrow v$ through channel l , and presents the interval from which the mean value of the distribution of channel gain is selected at each trial. Moreover, we assume $P_1 = 1$, $P_2 = 5$ and $\alpha = 10^{-3}$. Except for their instantaneous rewards, no other information is revealed to users. This information can be provided by the receiver feedback to transmitter. With these settings, it is easy to see that $((C_1, P_2), (C_2, P_2))$ is the unique pure strategy Nash equilibrium of this game, i.e. the theoretical convergence point.

2) *Results and Discussion*: We investigate the performance of selection strategies NR-BEWAS, NR-BFPLS and BERTS. The following strategies are also considered as benchmark:

- optimal (centralized) action (channel and power level) assignment that is based on global statistical channel knowledge and is performed by a central unit.
- uniformly random selection.

Figure 1 compares the average reward achieved by NR-BEWAS and NR-BFPLS by those of random and optimal selections. From the figure, despite being provided with only strictly limited information, both NR-BFPLS and NR-BEWAS exhibit vanishing regret, in the sense that the achieved average reward converges to that of centralized scenario.

⁷Note that this assumption is made in order to simplify the implementation; as established theoretically, all proposed procedures converge to equilibrium for arbitrary varying distributions.

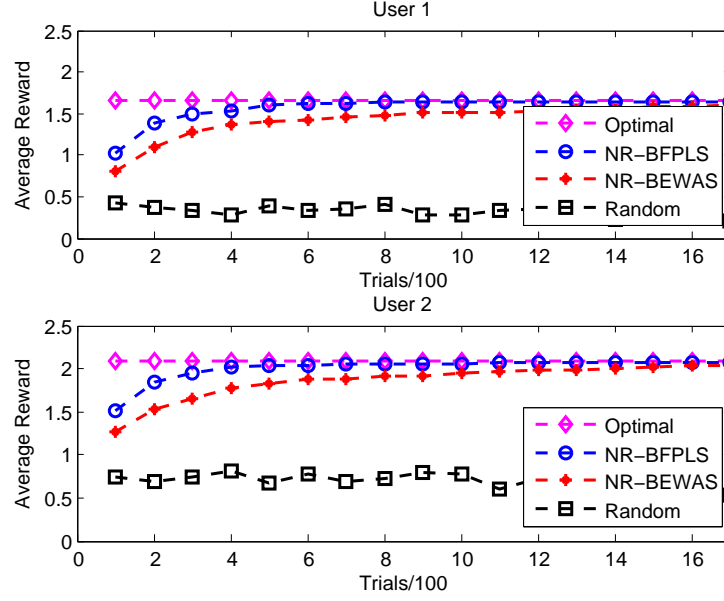


Fig. 1. Performance of four selection strategies. Both NR-BEWAS and NR-BFPLS exhibit vanishing regret; that is, their average rewards converge to that of optimal (centralized) selection.

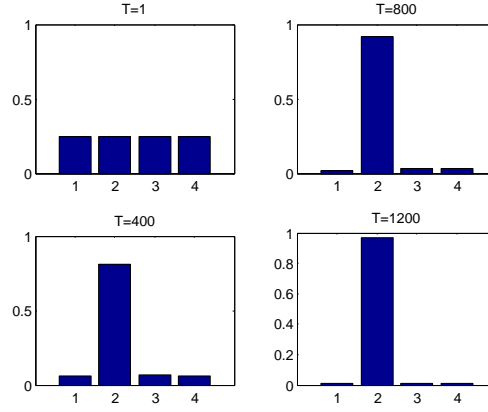


Fig. 2. Evolution of the mixed strategy of User 1, applying NR-BEWAS. Horizontal axis denotes the action indices, where index i , $i \in \{1, 2, 3, 4\}$, stands for action a_i . Vertical axis shows the weight of each action in the mixed strategy, i.e. its probability of being selected. The mixed strategy of User 1 converges to $\pi_1 = (0, 1, 0, 0)$.

Figures 2 and 3 illustrate the evolution of mixed strategies of the two users when NR-BEWAS is used. Figures 4 and 5, on the other hand, show the same variable when actions are selected by using NR-BFPLS. For both cases, the first and second users respectively converge to $a_2 : (C_1, P_2)$ and $a_4 : (C_2, P_2)$, as suggested by the theory.

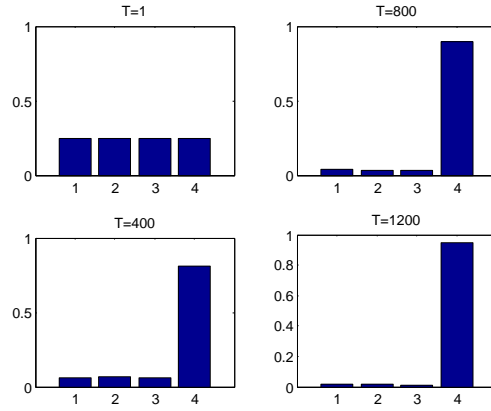


Fig. 3. Evolution of the mixed strategy of User 2, applying NR-BEWAS. The horizontal and vertical axes respectively depict the indices of actions and their selection probabilities. The mixed strategy of User 2 converges to $\pi_2 = (0, 0, 0, 1)$.

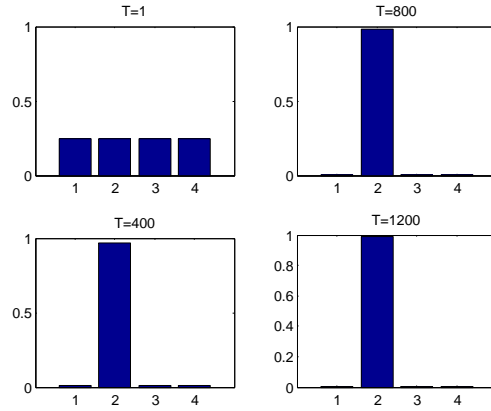


Fig. 4. Evolution of the mixed strategy of User 1, applying NR-BFPLS. The horizontal and vertical axes respectively depict the indices of actions and their selection probabilities. The mixed strategy of User 1 converges to $\pi_1 = (0, 1, 0, 0)$.

The performance of BERTS, however, is not an explicit function of game duration. As described before, the procedure continues to search mixed strategies until a suitable one, which yields a regret less than the selected threshold, is captured. Then this strategy is played for the rest of the game. Theorem 7.8 of [33] specifies the minimum game duration to guarantee the convergence of BERTS, which is relatively long even for small number of users and actions. Nevertheless, similar to other search-based algorithms, there also exists the possibility of finding some acceptable strategy at early stages of the game. As a result, for relatively short games, the

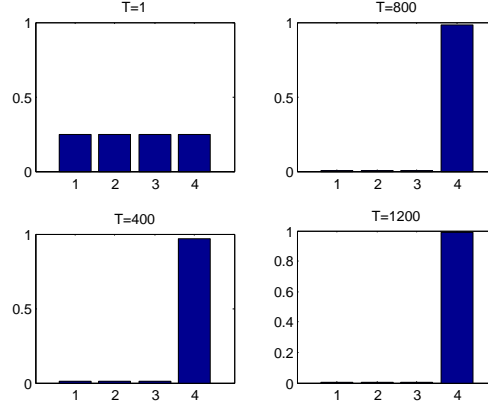


Fig. 5. Evolution of the mixed strategy of User 2, applying NR-BFPLS. The horizontal and vertical axes respectively depict the indices of actions and their selection probabilities. The mixed strategy of User 2 converges to $\pi_2 = (0, 0, 0, 1)$.

performance of BERTS is rather unpredictable. The other issue is the effect of regret threshold. On the one hand, larger threshold reduces the search time, since the set of acceptable strategies is large. On the other hand, large regret threshold might lead to performance loss, since there is the possibility that the user gets locked at some sub-optimal strategy at early stages, thereby incurring large accumulated regret. It is worth noting that due to its simplicity, and despite unpredictable performance, BERTS is an appealing approach in cases where computational effort should be minimized, and convergence to Nash equilibrium is desired. Figure 6 summarizes the results of few exemplary performances of BERTS. The parameters are selected as $T = 80$, $M = 1500$ and $\rho = 0.16$ (see Section VI). Simulation is performed for six *independent* rounds. The curve on the left side of Figure 6 depicts the period ($1 \leq m \leq 1500$) at which the algorithm finds an acceptable strategy. As expected, the results exhibit no specific pattern. The four sub-figures on the right depict the mixed strategies selected by BERTS at rounds 1 and 2, together with average rewards. From this figure, at round 2, acceptable strategies are found earlier than round 1 by both users, leading to better average performance. It is also worth noting that for User 2, the strategy of round 1 is in essence better than that of round 2; nevertheless, it is found later. As a result, the average performance of round 2 is superior to that of round 1.

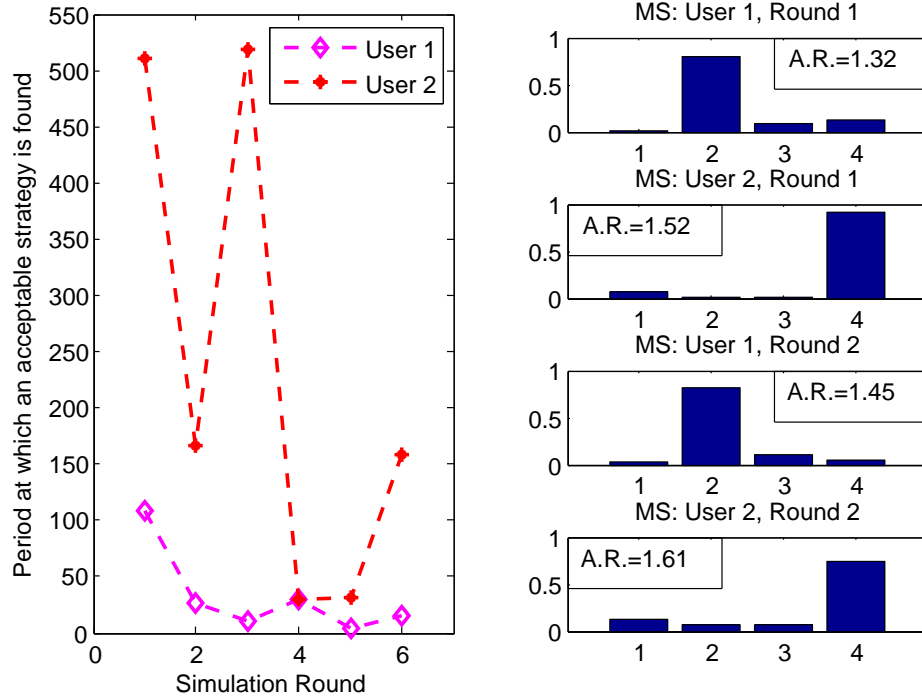


Fig. 6. Performance of BERTS. On the left, the vertical and horizontal axes show the periods and round number, respectively. The two curves depict the period at which a suitable mixed strategy (MS) is found at each of the 6 rounds. On the right, these mixed strategies are shown for both users at rounds 1 and 2, together with average rewards. The horizontal and vertical axes respectively depict the indices of actions and their selection probabilities.

B. Part Two

In this section we consider a wireless network consisting of 5 users (transmitter-receiver pairs), that compete for access to three orthogonal channels at two possible power levels (hence six actions). We compare BFPLS and BEWAS with the following selection approaches.⁸

- Optimal (centralized) action assignment as described in Section VII-A2.
- Centralized no-collision action selection, where no reward is assigned to users that access the same channel. Thus, users are encouraged to avoid collisions (a collision-avoidance

⁸As mentioned before, observing the joint action profile and/or communication among users is not required for implementing BEWAS, BFPLS and BERTS. Therefore, they cannot be compared with strategies that include mutual observation and/or communication. A good example of such algorithms is the widely-used *best-response dynamics*, where the strategy of each player is to play with the best-response to either the historical [10] or the predicted [5] joint action profile of opponents. Another example is the strategy suggested in [20], which is a combination of learning and auction algorithms where users communicate with each other.

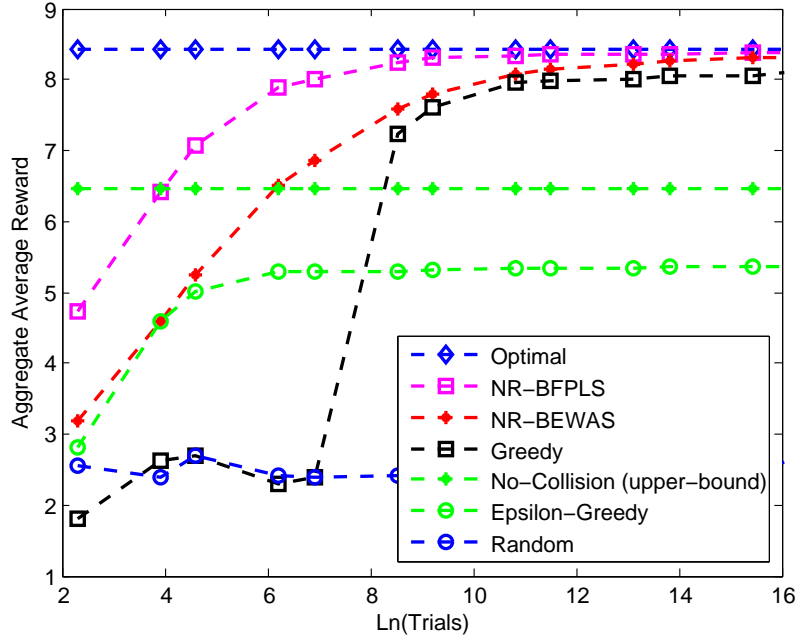


Fig. 7. Aggregate average reward of BFPLS and BEWAS compared to some other selection strategies.

strategy). This curve can be considered as an upper-bound for the performance of learning algorithms that select actions based on collision avoidance, such as [20].

- ϵ -greedy algorithm, where at each trial, with probability ϵ (exploration parameter), an action is selected uniformly at random, while with probability $1 - \epsilon$ the best action so far is played. The average reward of selected action is updated after each play [42]. For stationary environments, ϵ is usually time-varying and converges to zero in the limit, while in adversarial cases, ϵ is preferred to remain fixed. Here we let $\epsilon = 0.1$.
- Greedy approach, where at the beginning of the game, some trials are reserved for exploration, in which actions are selected at random (exploration period). The length of this period is a pre-defined fraction of the entire game duration. Based on the rewards of exploration period, the best possible action is selected, and is played for the rest of the game (exploitation period) [33]. This approach is extremely simple to implement; however, to the best of our knowledge, there is no analysis on the optimal length of the exploration period.
- Uniformly random selection.

The numerical results are depicted in Figure 7. From this figure, we can conclude the following.

- The performance of interference-avoidance strategies is strongly influenced by channel matrices and tends to be poor specifically when the number of channels is less than that of users. The reason is that the sum reward of multiple interfering users with limited transmit power might be larger than the maximum achievable reward of any single user.
- The performance of both BFPLS and BEWAS converge to that of centralized approach. As expected, BFPLS converges faster than BEWAS and we point out that the convergence speed of both algorithms would be dramatically enhanced if some side information was available to players, e.g. if users observed the actions of each other, or if communication was allowed among players. It is also worth noting that although BFPLS converges faster than BEWAS, the computation of integral (20) might be involved, especially for large number of actions [41].
- In general, ϵ -greedy and greedy approaches can be implemented easily with low computational cost; nevertheless, it can be seen that the greedy approaches are inferior to BEWAS and BFPLS in terms of asymptotic performance. Basically, these approaches are more suitable for stationary environments.

VIII. CONCLUSION AND REMARKS

This paper deals with resource allocation in multi-user infrastructureless wireless networks. The problem of utility maximization has been formulated using the multi-player multi-armed bandit theory framework. More precisely, given no side information, the users aim at minimizing some regret expressed in terms of the loss of reward by selecting appropriate actions on a given space of transmit power levels and orthogonal frequency channels. Based on some recent mathematical results, we have designed two selection strategies, which not only provide vanishing regret for each player, but also guarantee the asymptotic convergence of the game to the set of correlated equilibria. We have also studied experimental regret testing strategy that asymptotically converges to the set of Nash equilibria. Numerical results confirms the applicability of the game model and proposed strategies to wireless channel selection and power control.

IX. APPENDIX

A. Some Auxiliary Results

In this section, we state some auxiliary results and materials from game theory as well as bandit theory that are necessary for proofs.

1) *Game Theory*: Throughout this part, we consider a game \mathfrak{G} consisting of a set of K players where the strategy set of each player $k \in \{1, \dots, K\}$ is denoted by $I^{(k)}$ with a generic element $i^{(k)} = (i_1^{(k)}, \dots, i_M^{(k)})$. Similarly, the set of joint strategy profiles of players is denoted by \mathbf{I} with a generic element $\mathbf{i} = (i^{(1)}, \dots, i^{(K)})$ and \mathbf{i}_k^- stands for the joint action profile of all players except for player k . Moreover, $g^{(k)}(\mathbf{i})$ stands for the utility function of some player k .⁹

Definition 4. A game \mathfrak{G} is smooth if, for each $k \in \{1, \dots, K\}$, $g^{(k)}(\mathbf{i})$ has continuous partial derivatives with respect to the components of $i^{(k)}$.

Definition 5. Let $\nabla g^{(k)} = \left(\frac{\partial g^{(k)}}{\partial i_1^{(k)}}, \dots, \frac{\partial g^{(k)}}{\partial i_M^{(k)}} \right)$, and call $(\nabla g^{(k)})_{k \in \{1, \dots, K\}}$ the payoff gradient of a smooth game \mathfrak{G} . We say that the payoff gradient is strictly monotone if

$$\sum_{k=1}^K (\nabla g^{(k)}(\mathbf{i}) - \nabla g^{(k)}(\mathbf{j}))^T (i^{(k)} - j^{(k)}) < 0, \quad (23)$$

holds for all $\mathbf{i}, \mathbf{j} \in \mathbf{I}$ with $\mathbf{i} \neq \mathbf{j}$.

Theorem 2 ([43]). Consider a smooth game \mathfrak{G} with compact strategy sets. If the payoff gradient of \mathfrak{G} is strictly monotone then it has a unique correlated equilibrium, which places probability one on a unique pure-strategy Nash equilibrium.

Definition 6. A game \mathfrak{G} is potential if there exists a potential function $f : \mathbf{I} \rightarrow \mathbb{R}$ such that

$$g^{(k)}(i, \mathbf{i}_k^-) - g^{(k)}(j, \mathbf{i}_k^-) = f(i, \mathbf{i}_k^-) - f(j, \mathbf{i}_k^-), \quad (24)$$

for all $i, j \in I^{(k)}$ and $k \in \{1, \dots, K\}$.

Theorem 3 ([44]). Let \mathfrak{G} be a smooth potential game with a strictly concave potential function. Then a strategy profile is the unique pure strategy Nash equilibrium if and only if it is the potential maximizer.

⁹Note that compared to the system model some notation has been changed slightly.

Lemma 1 ([43]). *Let \mathfrak{G} be a smooth potential game. A potential of \mathfrak{G} is strictly concave if and only if the payoff gradient of \mathfrak{G} is strictly monotone.*

2) *Bandit Theory:*

Lemma 2. *Let R_n and R_{Ext} be given by (1) and (2), respectively. Then, for any $\delta \in (0, \frac{1}{2}]$, we have¹⁰*

$$\Pr \left(|R_n - R_{\text{Ext}}| \leq \sqrt{\frac{n}{2} \ln \frac{1}{\delta}} \right) \geq 1 - 2\delta, \quad (25)$$

from which it follows that if $R_n \in o(n)$, then we have $R_{\text{Ext}} \in o(n)$, with arbitrarily high probability.¹¹

Proof: By comparing (1) and (2), it suffices to show that $\Pr(|\sum_{t=1}^n (g_t(I_t) - \bar{g}_t(\mathbf{P}_t))|) \leq \sqrt{\frac{n}{2} \ln \frac{1}{\delta}} \leq 1 - 2\delta$. To this end, define $S := \sum_{t=1}^n g_t(I_t)$, where $g_t(I_t) \in [0, 1]$, $1 \leq t \leq n$, are independent random variables (see also Section II-A). Further note that $\bar{S} = \mathbb{E}[S] = \sum_{t=1}^n \bar{g}_t(\mathbf{P}_t)$. Therefore, by Hoeffding's inequality [33],

$$\begin{aligned} \Pr \left(|R_n - R_{\text{Ext}}| \geq \sqrt{\frac{n}{2} \ln \frac{1}{\delta}} \right) &= \Pr \left(|S - \bar{S}| \geq \sqrt{\frac{n}{2} \ln \frac{1}{\delta}} \right) \\ &\leq 2 \exp \left(-\frac{2 \frac{n}{2} \ln \frac{1}{\delta}}{n} \right) = 2\delta. \end{aligned} \quad (26)$$

Hence the Lemma follows with $\Pr \left(|R_n - R_{\text{Ext}}| \leq \sqrt{\frac{n}{2} \ln \frac{1}{\delta}} \right) = 1 - \Pr \left(|R_n - R_{\text{Ext}}| \geq \sqrt{\frac{n}{2} \ln \frac{1}{\delta}} \right)$. ■

Lemma 3. *Let R_{Ext} be given by (2). Moreover, define $\tilde{R}_n = \max_{i=1, \dots, N} \sum_{t=1}^n g_t(i) - \sum_{t=1}^n \tilde{g}_t(\mathbf{P}_t)$, where $\tilde{g}_t(\mathbf{P}_t) = \sum_{i=1}^N p_{i,t} \tilde{g}_t(i)$ and $\tilde{g}_t(i)$ is given by (12). Then we have*

$$\Pr \left(|\tilde{R}_n - R_{\text{Ext}}| \leq \sqrt{\frac{n}{2} \ln \frac{1}{\delta}} \right) \geq 1 - 2\delta. \quad (27)$$

Hence, for sufficiently small $\delta > 0$, $R_{\text{Ext}} \in o(n)$ implies that $\tilde{R}_n \in o(n)$, with arbitrarily high probability.

¹⁰Throughout this section and in order to simplify the notation, the player index (k) is omitted unless ambiguity arises.

¹¹Here and hereafter, the statement " $X(n) \in o(n)$ with arbitrarily high probability" for some nonnegative random sequence $X(n) \in \mathbb{R}$ means that the probability of $X(n) \notin o(n)$ can be made arbitrarily small, provided that some parameter is chosen sufficiently small.

Proof: Similar to the proof of Lemma 2, it follows from (2) and the definition of \tilde{R}_n that it is sufficient to show that $\Pr\left(|\sum_{t=1}^n (\tilde{g}_t(\mathbf{P}_t) - \bar{g}_t(\mathbf{P}_t))| \leq \sqrt{\frac{n}{2} \ln \frac{1}{\delta}}\right) \leq 1 - 2\delta$ for $\delta \in (0, \frac{1}{2}]$. To this end, note that $\tilde{g}_t(\mathbf{P}_t) \in [0, 1]$, $1 \leq t \leq n$, are independent random variables. Moreover, since $\tilde{g}_t(i)$ is an unbiased estimate of $g_t(i)$, we have $\mathbb{E}[\sum_{t=1}^n \tilde{g}_t(\mathbf{P}_t)] = \sum_{t=1}^n \bar{g}_t(\mathbf{P}_t)$. Hence, defining $S = \sum_{t=1}^n (\tilde{g}_t(\mathbf{P}_t) - \bar{g}_t(\mathbf{P}_t))$ and proceeding as in the proof of Lemma 2 with the Hoeffding's inequality in hand proves the lemma. ■

Proposition 6. *Let R_n be given by (1) and \tilde{R}_n be defined as in Lemma 3. Then, $R_n \in o(n)$ implies that $\tilde{R}_n \in o(n)$.*

Proof: Lemma 2 implies that $R_n \in o(n) \Rightarrow R_{\text{Ext}} \in o(n)$ with arbitrarily high probability, while by Lemma 3, we have $R_{\text{Ext}} \in o(n) \Rightarrow \tilde{R} \in o(n)$. Therefore, if $R_n \in o(n)$, then $\tilde{R} \in o(n)$ with arbitrarily high probability. ■

Theorem 4. ([33]) *Let $\Phi(\mathbf{U}) = \psi(\sum_{i=1}^N \phi(u_i))$, where $\mathbf{U} = (u_1, \dots, u_N)$. Consider a selection strategy, which at time t selects action I_t according to distribution \mathbf{P}_t , whose elements $p_{i,t}$ are defined as*

$$p_{i,t} = (1 - \gamma_t) \frac{\phi'(R_{i,t-1})}{\sum_{k=1}^N \phi'(R_{k,t-1})} + \frac{\gamma_t}{N}, \quad (28)$$

where $R_{i,t-1} = \sum_{s=1}^{t-1} (g_s(i) - g_s(I_s))$. Assume that:

A1. $\sum_{t=1}^n \frac{1}{\gamma_t^2} = o(\frac{n^2}{\ln n})$,

A2. For all vectors $\mathbf{V}_t = (v_{1,t}, \dots, v_{n,t})$ with $|v_{i,t}| \leq \frac{N}{\gamma_t}$, we have

$$\lim_{n \rightarrow \infty} \frac{1}{\psi(\phi(n))} \sum_{t=1}^n C(\mathbf{V}_t) = 0, \quad (29)$$

where $C(\mathbf{V}_t) = \sup_{\mathbf{U} \in \mathbb{R}^N} \psi'(\sum_{i=1}^N \phi(u_i)) \sum_{i=1}^N \phi''(u_i) v_{i,t}^2$.

A3. For all vectors $\mathbf{U}_t = (u_{1,t}, \dots, u_{n,t})$, with $u_{i,t} \leq t$,

$$\lim_{n \rightarrow \infty} \frac{1}{\psi(\phi(n))} \sum_{t=1}^n \gamma_t \sum_{i=1}^N \nabla_i \Phi(\mathbf{U}_t) = 0. \quad (30)$$

A4. For all vectors $\mathbf{U}_t = (u_{1,t}, \dots, u_{n,t})$, with $u_{i,t} \leq t$,

$$\lim_{n \rightarrow \infty} \frac{\ln n}{\psi(\phi(n))} \sqrt{\sum_{t=1}^n \frac{1}{\gamma_t^2} \left(\sum_{i=1}^N \nabla_i \Phi(\mathbf{U}_t) \right)^2}. \quad (31)$$

Then the selection strategy satisfies

$$\lim_{n \rightarrow \infty} \frac{1}{n} \left(\max_{i=1, \dots, N} \sum_{t=1}^n g_t(i) - \sum_{t=1}^n g_t(I_t) \right) = 0, \quad (32)$$

or equivalently, $R_n \in o(n)$, where R_n is given by (1).

B. Proof of Proposition 1

In order to prove Proposition 1, we use Theorem 2. As the strategy set is compact, in order to use this theorem, we show that 1) the game is smooth, and 2) the payoff gradient is strictly monotone.

According to our system model, by changing the channel index, $c^{(k)}$, the channel gain and interference changes. Therefore we define $i_1^{(k)} := \frac{|h_{kk'}, c^{(k)}|^2}{\sum_{q=1}^{Q_k} l^{(q)} |h_{qk'}, c^{(k)}|^2 + N_0}$ and $i_2^{(k)} := l^{(k)}$, from which we have $g^{(k)}(\mathbf{i}) = \log(i_1^{(k)} i_2^{(k)}) - \alpha i_2^{(k)}$. This results in

$$\frac{\partial g^{(k)}}{\partial i_1^{(k)}} = \frac{1}{i_1^{(k)}}. \quad (33)$$

and

$$\frac{\partial g^{(k)}}{\partial i_2^{(k)}} = \frac{1}{i_2^{(k)}} - \alpha. \quad (34)$$

Hence by Definition 4, the game is smooth. On the other hand, given $g^{(k)}$, we have

$$\begin{aligned} & (\nabla g^{(k)}(\mathbf{i}) - \nabla g^{(k)}(\mathbf{j}))^T (i^{(k)} - j^{(k)}) \\ &= \begin{bmatrix} \frac{1}{i_1^{(k)}} - \frac{1}{j_1^{(k)}} & \frac{1}{i_2^{(k)}} - \frac{1}{j_2^{(k)}} \end{bmatrix} \begin{bmatrix} i_1^{(k)} - j_1^{(k)} \\ i_2^{(k)} - j_2^{(k)} \end{bmatrix} \\ &= \left(\frac{1}{i_1^{(k)}} - \frac{1}{j_1^{(k)}} \right) (i_1^{(k)} - j_1^{(k)}) + \left(\frac{1}{i_2^{(k)}} - \frac{1}{j_2^{(k)}} \right) (i_2^{(k)} - j_2^{(k)}), \end{aligned} \quad (35)$$

which is always negative as for any $x, y > 0$ and $x \neq y$, $x - y > 0$ yields $\frac{1}{x} - \frac{1}{y} < 0$ and vice versa. So,

$$\sum_{k=1}^K \nabla g^{(k)} < 0 \quad (36)$$

i.e. the payoff gradient is strictly monotone by Definition 5.

As a result, by Theorem 2, the game has a unique correlated equilibrium which places

probability one on the unique Nash equilibrium.

C. Proof of Proposition 2

First, we point out that the game is a potential game with a potential function being $f(\cdot) = \sum_{k=1}^K g^{(k)}(\cdot)$, by simply inserting $g^{(k)}$ and f in condition (24). Moreover, similar to Proposition 1, it can be easily shown that the game is smooth and the payoff gradient is strictly monotone (Define $i_1^{(k)} := \frac{|h_c(k)|^2}{N_0}$ and $i_2^{(k)} = l^{(k)}$).¹² Therefore, by Lemma 1, the game is a smooth potential game with strictly concave potential function. As a result, by Theorem 3, it has a unique pure strategy Nash equilibrium which is the potential maximizer. On the other hand, by Lemma 2, the game has a unique correlated equilibrium which places probability one on the unique pure strategy Nash equilibrium.

D. Proof of Proposition 3

We first notice that $R_{\text{Ext}} \in O((nN)^{\frac{2}{3}}(\ln N)^{\frac{1}{3}})$, as stated by the following lemma.¹³

Lemma 4. *Consider a selection strategy that uses $\mathbf{P}_t = (p_{1,t}, \dots, p_{N,t})$ to select an action among N possible choices, where $p_{i,t}$ is calculated as*

$$p_{i,t} = (1 - \gamma) \frac{\exp(\eta \tilde{R}_{i,t-1})}{\sum_{m=1, \dots, N} \exp(\eta \tilde{R}_{m,t-1})} + \frac{\gamma}{N}, \quad (37)$$

and $\tilde{R}_{i,t-1}$ denotes the estimated accumulated regret of not playing action i .¹⁴ Then selecting γ and η as given by Proposition 3 yields $R_{\text{Ext}} \in O((nN)^{\frac{2}{3}}(\ln N)^{\frac{1}{3}})$.

Proof: The proof is a direct corollary of Theorem 6.6 of [33]. ■

Given Lemma 4, we follow the approach of [34] for the rest of the proof.

Recall that by Section II-C and Algorithm 1, the mixed strategy of each player is defined by

$$\mathbf{P}_t = \sum_{(i \rightarrow j): i \neq j} \mathbf{P}_t^{(i \rightarrow j)} \delta_{(i \rightarrow j), t}. \quad (38)$$

¹²Details are similar to Proposition 1 and thus are omitted.

¹³Throughout this section and in order to simplify the notation, the player index (k) is omitted unless ambiguity arises.

¹⁴This definition should not be mistaken for the general regret defined in Section II-A.

Hence,

$$\bar{g}_t(\mathbf{P}_t) = \sum_{(i \rightarrow j): i \neq j} \bar{g}_t(\mathbf{P}_t^{(i \rightarrow j)}) \delta_{(i \rightarrow j), t}. \quad (39)$$

Lemma 4 specifies the growth rate of external regret. On the other hand, as described in Section II-C, the convergence approach applies the BEWAS algorithm for $N(N-1) \leq N^2$ actions. Therefore, (39) together with Lemma 4 yields

$$\max \sum_{t=1}^n \bar{g}_t(\mathbf{P}_t^{(i \rightarrow j)}) - \sum_{t=1}^n \bar{g}_t(\mathbf{P}_t) \in O((N^2 n)^{\frac{2}{3}} (2 \ln N)^{\frac{1}{3}}), \quad (40)$$

and the definition of internal regret ensures that $\max_{i \neq j} R_{(i \rightarrow j), n} \in O((N^2 n)^{\frac{2}{3}} (\ln N)^{\frac{1}{3}})$, which concludes the proof. Details can be found in [34], and hence are omitted.

E. Proof of Proposition 4

We first show that the algorithm has vanishing external regret, i.e. $R_{\text{Ext}} \in o(n)$, as formalized in the following.

Lemma 5. *Consider a selection strategy that uses $\mathbf{P}_t = (p_{1,t}, \dots, p_{N,t})$ to select an action among N possible choices, where $p_{i,t}$ is calculated as*

$$p_{i,t} = (1 - \gamma_t) \frac{\exp(\eta_t \tilde{R}_{i,t-1})}{\sum_{m=1, \dots, N} \exp(\eta_t \tilde{R}_{m,t-1})} + \frac{\gamma_t}{N}, \quad (41)$$

and $\tilde{R}_{i,t-1}$ denotes the estimated accumulated regret of not playing action i . Then, for γ_t and η_t as given by Proposition 4, this strategy yields vanishing external regret, i.e. $R_{\text{Ext}} \in o(n)$.

Proof: By Proposition 6, if (32) is satisfied for a selection strategy (that is, if $R_n \in o(n)$), then the growth rate of the external regret caused by the bandit version of that strategy (which uses estimated rewards instead of true ones) grows sublinearly in n , i.e. $\tilde{R}_n \in o(n)$. Therefore, in order to prove the proposition, we can show that our selected parameters $\gamma_t = t^{-\frac{1}{3}}$ and $\eta_t = \frac{\gamma_t^3}{N^2}$ satisfy axioms A1-A4 of Theorem 4. In what follows, we show that each of these axioms is fulfilled. In doing so, we omit for the lack of space simple calculus steps. Also the reader should note that in our strategy we have $\Phi(\mathbf{U}) = \frac{1}{\eta_t} \ln \left(\sum_{i=1}^N \exp(\eta_t u_i) \right)$.

A1. For $\gamma_t = t^{-\frac{1}{3}}$, we have

$$\sum_{t=1}^n \frac{1}{\gamma_t^2} = \sum_{t=1}^n t^{\frac{2}{3}} = \text{Harmonic Number}[n, -\frac{2}{3}] := H_n[-\frac{2}{3}]. \quad (42)$$

Then,

$$\lim_{n \rightarrow \infty} \frac{\ln n}{n^2} \sum_{t=1}^n \gamma_t^2 = \lim_{n \rightarrow \infty} \frac{\ln n}{n^2} H_n[-\frac{2}{3}] = 0. \quad (43)$$

A2. For $\psi(x) = \frac{1}{\eta_t} \ln x$ and $\phi(x) = \exp(\eta_t x)$, we obtain

$$C(\mathbf{V}_t) = \sup \left(\eta_t \sum_{i=1}^N v_{i,t}^2 \right) = \frac{\eta_t N^3}{\gamma_t^2}. \quad (44)$$

Hence,

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{\psi(\phi(n))} \sum_{t=1}^n C(\mathbf{V}_t) &= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n t^{\frac{-1}{3}} \\ &= \lim_{n \rightarrow \infty} \frac{1}{n} H_n[\frac{1}{3}] = 0. \end{aligned} \quad (45)$$

A3. For $\Phi(\mathbf{U}) = \frac{1}{\eta_t} \ln \left(\sum_{i=1}^N \exp(\eta_t u_i) \right)$, $\nabla_i \Phi(\mathbf{U}_t)$ yields

$$\nabla_i \Phi(\mathbf{U}_t) = \frac{\exp(\eta_t u_i)}{\sum_{i=1}^N \exp(\eta_t u_i)}. \quad (46)$$

Therefore,

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{\psi(\phi(n))} \sum_{t=1}^n \gamma_t \sum_{i=1}^N \nabla_i \Phi(\mathbf{U}_t) &= \\ \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n t^{\frac{-1}{3}} \sum_{i=1}^N \frac{\exp(\eta_t u_i)}{\sum_{i=1}^N \exp(\eta_t u_i)} &= \\ \lim_{n \rightarrow \infty} \frac{1}{n} H_n[\frac{1}{3}] &= 0. \end{aligned} \quad (47)$$

A4. A4 follows simply by substituting (46) in (31).

Hence, all axioms A1-A4 are satisfied, and therefore (32) holds, which, together with Proposition 6, completes the proof. ■

By Lemma 5, the external regret BEWAS grows sublinearly in n . Therefore, similar to the proof of Proposition 3, (39) yields

$$\max \sum_{t=1}^n \bar{g}_t(\mathbf{P}_t^{(i \rightarrow j)}) - \sum_{t=1}^n \bar{g}_t(\mathbf{P}_t) \in o(n), \quad (48)$$

and the definition of internal regret ensures that $\max_{i \neq j} R_{(i \rightarrow j), n} \in o(n)$, which concludes the proof.

REFERENCES

- [1] H. Robbins, "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematical Society*, vol. 58, no. 5, pp. 527–535, 1952.
- [2] K. Liu, Q. Zhao, and B. Krishnamachari, "Distributed learning under imperfect sensing in cognitive radio networks," in *Asilomar Conference on Signals, Systems and Computers*, Nov 2010, pp. 671–675.
- [3] K. Liu and Q. Zhao, "Cooperative game in dynamic spectrum access with unknown model and imperfect sensing," *IEEE Transactions on Wireless Communications*, vol. 11, no. 4, pp. 1596–1604, April 2012.
- [4] M. Di Felice, K.R. Chowdhury, and L. Bononi, "Learning with the bandit: A cooperative spectrum selection scheme for cognitive radio networks," in *IEEE Global Telecommunications Conference*, Dec 2011, pp. 1–6.
- [5] S. Maghsudi and S. Stanczak, "Relay selection problem with no side information: An adversarial bandit approach," in *IEEE Wireless Communications and Networking Conference*, April 2013, pp. 715–720.
- [6] V. Krishnamurthy and D.V. Djonin, "Structured threshold policies for dynamic sensor scheduling—a partially observed Markov decision process approach," *IEEE Transactions on Signal Processing*, vol. 55, no. 10, pp. 4938–4957, Oct 2007.
- [7] J. Nino-Mora and S.S. Villar, "Sensor scheduling for hunting elusive hiding targets via Whittle's restless bandit index policy," in *International Conference on Network Games, Control and Optimization*, Oct 2011, pp. 1–8.
- [8] P. Arora, C. Szepesvari, and R. Zheng, "Sequential learning for optimal monitoring of multi-channel wireless networks," in *IEEE International Conference on Computer Communications*, April 2011, pp. 1152–1160.
- [9] R. Zheng, T. Le, and Z. Han, "Approximate online learning for passive monitoring of multi-channel wireless networks," in *IEEE International Conference on Computer Communications*, April 2013, pp. 3111–3119.
- [10] T. Chen, L. Zhu, F. Wu, and S. Zhong, "Stimulating cooperation in vehicular ad hoc networks: A coalitional game theoretic approach," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 2, pp. 566–579, Feb 2011.
- [11] W. Saad, Z. Han, T. Basar, M. Debbah, and A. Hjørungnes, "Hedonic coalition formation for distributed task allocation among wireless agents," *IEEE Transactions on Mobile Computing*, vol. 10, no. 9, pp. 1327–1344, Sept 2011.
- [12] A. Mukherjee and H.M. Kwon, "General auction-theoretic strategies for distributed partner selection in cooperative wireless networks," *IEEE Transactions on Communications*, vol. 58, no. 10, pp. 2903–2915, Oct 2010.
- [13] J. Sun, E. Modiano, and L. Zheng, "Wireless channel allocation using an auction algorithm," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 5, pp. 1085–1096, May 2006.
- [14] O. Ileri, M. Siun-Chuon, and N.B. Mandayam, "Pricing for enabling forwarding in self-configuring ad hoc networks," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 1, pp. 151–162, Jan 2005.
- [15] S. Maghsudi and S. Stanczak, "A hybrid centralized-decentralized resource allocation scheme for two-hop transmission," in *International Symposium on Wireless Communication Systems*, Nov 2011, pp. 96–100.
- [16] M. Guo, Y. Liu, and J. Malec, "A new Q-learning algorithm based on the metropolis criterion," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 34, no. 5, pp. 2140–2143, Oct 2004.
- [17] X. Fang, D. Yang, and G. Xue, "Taming wheel of fortune in the air: An algorithmic framework for channel selection strategy in cognitive radio networks," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 2, pp. 783–796, Feb 2013.
- [18] Y. Song, Y. Fang, and Y. Zhang, "Stochastic channel selection in cognitive radio networks," in *IEEE Global Telecommunications Conference*, Nov 2007, pp. 4878–4882.

- [19] Y. Gai, B. Krishnamachari, and R. Jain, "Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations," *IEEE/ACM Transactions on Networking*, vol. 20, no. 5, pp. 1466–1478, Oct 2012.
- [20] D. Kalathil, N. Nayyar, and R. Jain, "Decentralized learning for multi-player multi-armed bandits," in *IEEE Annual Conference on Decision and Control*, Dec 2012, pp. 3960–3965.
- [21] Y. Xu, J. Wang, Q. Wu, A. Anpalagan, and Y.D. Yao, "Opportunistic spectrum access in unknown dynamic environment: A game-theoretic stochastic learning solution," *IEEE Transactions on Wireless Communications*, vol. 11, no. 4, pp. 1380–1391, April 2012.
- [22] Y. Xu, Q. Wu, L. Shen, J. Wang, and A. Anpalagan, "Opportunistic spectrum access with spatial reuse: Graphical game and uncoupled learning solutions," *IEEE Transactions on Wireless Communications*, vol. 12, no. 10, pp. 4814–4826, Oct 2013.
- [23] Y. Xu, Q. Wu, J. Wang, L. Shen, and A. Anpalagan, "Opportunistic spectrum access using partially overlapping channels: Graphical game and uncoupled learning," *IEEE Transactions on Communications*, vol. 61, no. 9, pp. 3906–3918, Sep 2013.
- [24] A. Anandkumar, N. Michael, and A. Tang, "Opportunistic spectrum access with multiple users: Learning under competition," in *IEEE International Conference on Computer Communications*, March 2010, pp. 1–9.
- [25] W. Xu, L. Liang, H. Zhang, S. Jin, J.C.F. Li, and M. Lei, "Performance enhanced transmission in device-to-device communications: Beamforming or interference cancellation?," in *IEEE Global Communications Conference*, Dec 2012, pp. 4296–4301.
- [26] K. Liu, Q. Zhao, and B. Krishnamachari, "Decentralized multi-armed bandit with imperfect observations," in *Annual Allerton Conference on Communication, Control, and Computing*, Sept 2010, pp. 1669–1674.
- [27] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE Transactions on Signal Processing*, vol. 58, no. 11, pp. 5667–5681, Nov 2010.
- [28] H. Liu, K. Liu, and Q. Zhao, "Learning in a changing world: Restless multiarmed bandit with unknown dynamics," *IEEE Transactions on Information Theory*, vol. 59, no. 3, pp. 1902–1916, March 2013.
- [29] A. Blum and Y. Mansour, "From external to internal regret," *Journal of Machine Learning Research*, vol. 8, pp. 1307–1324, Dec 2007.
- [30] P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire, "The non-stochastic multi-armed bandit problem," *SIAM Journal on Computing*, vol. 32, no. 1, pp. 48–77, Jan 2003.
- [31] J. Kujala and T. Elomaa, "On following the perturbed leader in the bandit setting," in *Algorithmic Learning Theory*, Oct 2005, pp. 371–385.
- [32] F. Germano and G. Lugosi, "Global Nash convergence of Foster and Youngs regret testing," *Games and Economic Behavior*, vol. 60, no. 1, pp. 154, July 2007.
- [33] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*, Cambridge University Press, 2006.
- [34] G. Stoltz and G. Lugosi, "Internal regret in on-line portfolio selection," *Journal of Machine Learning*, vol. 59, no. 1, pp. 125–159, 2005.
- [35] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.
- [36] N. Nisan, T. Roughgarden, E. Tardos, and V. Vazirani, *Algorithmic Game Theory*, Cambridge University Press, 2007.

- [37] S. Hart and A. Mas-colell, "A general class of adaptive strategies," *Journal of Economic Theory*, vol. 98, pp. 26–54, May 2001.
- [38] N. Cesa-Bianchi and G. Lugosi, "Potential-based algorithms in on-line prediction and game theory," *Journal of Machine Learning*, vol. 51, no. 3, pp. 239–261, 2003.
- [39] Y.D. Yao and A.U.H. Sheikh, "Approximation to Bayes risk in repeated play," *Contributions to the Theory of Games*, vol. 3, no. 39, pp. 97–139, 1957.
- [40] J. Kujala and T. Elomaa, "Following the perturbed leader to gamble at multi-armed bandits," in *Algorithmic Learning Theory*, 2007, vol. 4754, pp. 166–180.
- [41] M. Hutter and J. Poland, "Adaptive online prediction by following the perturbed leader," *Journal of Machine Learning Research*, vol. 6, pp. 639–660, Dec 2005.
- [42] J. Nie and S. Haykin, "A Q-learning-based dynamic channel assignment technique for mobile communication systems," *IEEE Transactions on Vehicular Technology*, vol. 48, no. 5, pp. 1676–1687, Sep 1999.
- [43] T. Ui, "Correlated equilibrium and concave games," *International Journal of Game Theory*, vol. 37, no. 1, pp. 1–13, April 2008.
- [44] A. Nayman, "Correlated equilibrium and potential games," *International Journal of Game Theory*, vol. 26, no. 2, pp. 223–227, 1997.